

Fast sampling for Bayesian inference in neural circuits

Guillaume Hennequin^{*1}, Laurence Aitchison², and Máté Lengyel¹

¹Computational & Biological Learning Lab, Department of Engineering, University of Cambridge, UK

²Gatsby Computational Neuroscience Unit, University College London, UK

This manuscript is a preliminary written version of our Cosyne poster [1]

Abstract

Time is at a premium for recurrent network dynamics, and particularly so when they are stochastic and correlated: the quality of inference from such dynamics fundamentally depends on how fast the neural circuit generates new samples from its stationary distribution. Indeed, behavioral decisions can occur on fast time scales (~ 100 ms), but it is unclear what neural circuit dynamics afford sampling at such high rates. We analyzed a stochastic form of rate-based linear neuronal network dynamics with synaptic weight matrix \mathbf{W} , and the dependence on \mathbf{W} of the covariance of the stationary distribution of joint firing rates. This covariance Σ can be actively used to represent posterior uncertainty via sampling under a linear-Gaussian latent variable model. The key insight is that the mapping between \mathbf{W} and Σ is degenerate: there are infinitely many \mathbf{W} 's that lead to sampling from the same Σ but differ greatly in the speed at which they sample. We were able to explicitly separate these extra degrees of freedom in a parametric form and thus study their effects on sampling speed. We show that previous proposals for probabilistic sampling in neural circuits correspond to using a symmetric \mathbf{W} which violates Dale's law and results in critically slow sampling, even for moderate stationary correlations. In contrast, optimizing network dynamics for speed consistently yielded asymmetric \mathbf{W} 's and dynamics characterized by fast transients, such that samples of network activity became fully decorrelated over ~ 10 ms. Importantly, networks with separate excitatory/inhibitory populations proved to be particularly efficient samplers, and were in the balanced regime. Thus, plausible neural circuit dynamics can perform fast sampling for efficient decoding and inference.

1 Introduction

Perception in humans is blazingly fast: when presented with an image for 20 ms, we can tell in a split second whether or not it contained an animal, and our brain holds the correct answer as early as 150 ms following stimulus onset [2]. Such celerity is surprising given the difficulty of the task: sensory inputs being noisy and ambiguous (Figure 1A), they do not uniquely determine the state of the environment, so perception is inherently a matter of probabilistic inference [3]. Thus, the brain must represent and compute with complex probability distributions over relevant environmental variables. Most state-of-the-art machine learning techniques for solving similar inference problems at large scale face a tradeoff between inference accuracy and computing speed (e.g. [4]). The brain, on the contrary, seems to enjoy both simultaneously.

Some probabilistic computations can be made easier through an appropriate choice of representation for the probability distributions of interest. Sampling-based representations (Figure 1B, [5, 6]), for example, make computing moments of the distribution or its marginals

straightforward. However, the speed issue (cf. above) becomes even more fundamental: no matter how close the actual sampled distribution is to the ideal one, any sampling-based computation becomes accurate only after enough samples have been collected, and one has no choice but *waiting* for those samples to be delivered by the circuit dynamics. For sampling to be of any practical use, the interval that separates the generation of two independent samples must be short relative to the desired behavioural timescale (Figure 1C). Single neurons can integrate their inputs on a timescale $\tau_m \approx 10 - 50$ ms, whereas we must often make decisions in less than a second: this leaves just enough time to use (i.e. read out) a few tens of samples. Thus, it seems that the dynamics of brain circuits cannot afford correlating neural activity on a timescale longer than τ_m . How such temporal decorrelation can be achieved in cortical circuits remains unclear.

In this note, we introduce a simple yet non-trivial generative model and seek plausible neuronal network dynamics for *fast* sampling from the corresponding posterior distribution. While some standard machine learning techniques do suggest “neural network”-type solutions to sampling, not only are the corresponding architectures

^{*}gje2@cam.ac.uk

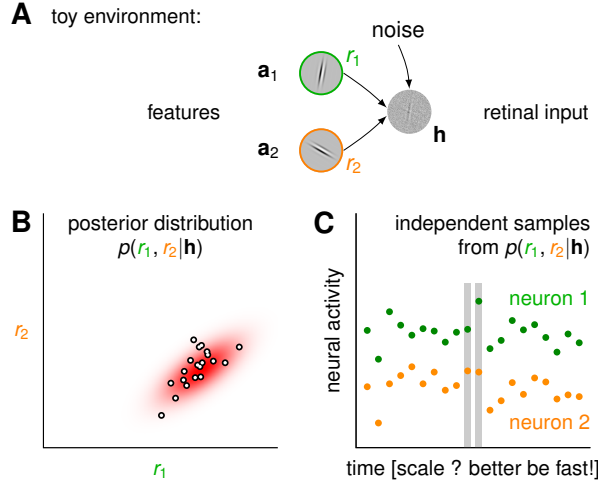


Figure 1: **Sampling-based representation of (perceptual) uncertainty.** A toy visual environment (A) comprises two features (oriented edges) \mathbf{a}_1 and \mathbf{a}_2 which are present in the scene with intensities r_1 and r_2 respectively. The two features combine linearly to form the “retinal input” \mathbf{h} , to which noise is added. Perception is about inferring the intensities r_1 and r_2 at which the features are present in the scene, given the retinal input \mathbf{h} . If the uncertainty about r_1 and r_2 matters, e.g. for making optimal decisions when they too have uncertain consequences [5], one must represent the full distribution $p(r_1, r_2|\mathbf{h})$ (B). This can be done by drawing (independent) samples from that distribution (B, white dots), which the brain could encode in the joint activity of two neurons (C; each time frame corresponds to one of the white dots in B). For sampling to be of any practical use, the minimum time that separates the collection of two independent samples (gray frames in C) must be short.

implausible in fundamental ways (e.g. they violate Dale’s law), but we show here that they lead to unacceptably slow sampling in high dimensions. Although this problem is already well appreciated in the machine learning community, the simplicity of our generative model allows us to draw an analytical picture of it and to suggest solutions. In fact, we can use methods from robust control to discover the *fastest* neural-like sampler for our generative model, and study its structure. We find that it corresponds to greatly non-symmetric synaptic interactions (in contrast to most off-the-shelf samplers), and mathematically nonnormal circuit dynamics [7], in striking agreement with our current understanding of primary visual cortex (V1) dynamics [8].

2 Basic setup

We focus on the linear Gaussian latent variable model, which is a high-dimensional generalization of the example given in Figure 1A. The model generates observations $\mathbf{h} \in$

\mathbb{R}^M as weighted sums of N features $(\mathbf{a}_1, \dots, \mathbf{a}_N) \in \mathbb{R}^{M \times N}$ with jointly Gaussian coefficients (r_1, \dots, r_N) , plus independent additive noise terms (Figure 2, left). More formally:

$$p(\mathbf{r}) = \mathcal{N}(\mathbf{r}; 0, \mathbf{C}) \quad (1)$$

$$p(\mathbf{h}|\mathbf{r}) = \mathcal{N}(\mathbf{h}; \mathbf{A}\mathbf{r}, \sigma_h^2 \mathbf{I}) \quad (2)$$

$$\mathbf{A} = (\mathbf{a}_1; \mathbf{a}_2; \dots; \mathbf{a}_N) \quad (3)$$

The posterior distribution is multivariate Gaussian:

$$p(\mathbf{r}|\mathbf{h}) = \mathcal{N}(\mathbf{r}; \boldsymbol{\mu}(\mathbf{h}), \boldsymbol{\Sigma}) \quad (4)$$

$$\boldsymbol{\Sigma} = \left(\mathbf{C}^{-1} + \frac{\mathbf{A}^\top \mathbf{A}}{\sigma_h^2} \right)^{-1} \quad (5)$$

$$\boldsymbol{\mu}(\mathbf{h}) = \boldsymbol{\Sigma} \mathbf{A}^\top \mathbf{h} / \sigma_h^2. \quad (6)$$

We are interested in neural circuit dynamics for sampling from $p(\mathbf{r}|\mathbf{h})$, whereby the data (observation) \mathbf{h} is given as a constant feedforward input to another layer of recurrently connected units, which encode the latent variables and also receive inputs from external, private sources of noise $\boldsymbol{\xi}$ (Figure 2, right). The activity fluctuations $\mathbf{r}(t)$ in the recurrent layer must have a stationary distribution that matches the posterior, for any \mathbf{h} .

More precisely, we consider linear recurrent stochastic dynamics of the form:

$$d\mathbf{r} = \frac{dt}{\tau_m} (-\mathbf{r}(t) + \mathbf{W}\mathbf{r}(t) + \mathbf{F}\mathbf{h}) + \sigma_\xi \sqrt{\frac{2}{\tau_m}} d\boldsymbol{\xi}(t) \quad (7)$$

Here τ_m is the single-unit “membrane” time constant, and $d\boldsymbol{\xi}$ is a Wiener process of unit variance, which is scaled by a scalar noise intensity σ_ξ . The activity $r_i(t)$ could represent either the membrane potential of neuron i , or the deviation from baseline of its momentary firing rate. The matrices \mathbf{F} and \mathbf{W} contain the feedforward and recurrent connection weights, respectively.

The stationary distribution of \mathbf{r} is indeed Gaussian with a mean

$$\boldsymbol{\mu}^r(\mathbf{h}) = (\mathbf{I} - \mathbf{W})^{-1} \mathbf{F}\mathbf{h} \quad (8)$$

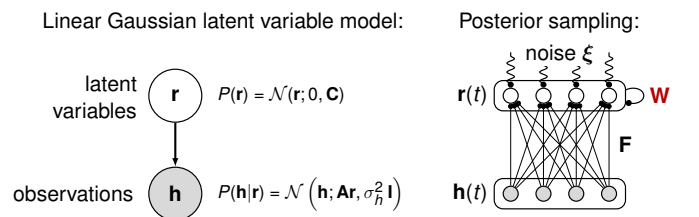


Figure 2: **Sampling under a linear Gaussian latent variable model using neuronal network dynamics.** Left: schematics of the generative model. Right: schematics of the recognition model. Sampling from $p(\mathbf{r}|\mathbf{h})$ is achieved through the linear, recurrent processing of both the input \mathbf{h} and some private sources of noise $\boldsymbol{\xi}$ (see text). \mathbf{F} and \mathbf{W} denote feedforward and recurrent synaptic weight matrices respectively.

and a covariance matrix $\Sigma^r \equiv \langle (\mathbf{r}(t) - \boldsymbol{\mu}^r)(\mathbf{r}(t) - \boldsymbol{\mu}^r)^\top \rangle_t$ that depends only on \mathbf{W} and σ_ξ , but not on \mathbf{h} , according to the following Lyapunov equation [9]:

$$(\mathbf{W} - \mathbf{I})\Sigma^r + \Sigma^r(\mathbf{W} - \mathbf{I})^\top = -2\sigma_\xi^2 \mathbf{I} \quad (9)$$

where \mathbf{I} denotes the identity matrix. Note that in the absence of a recurrent connectivity ($\mathbf{W} = 0$), the variance of every $r_i(t)$ would be exactly σ_ξ^2 .

In order for the dynamics of Equation (7) to sample from the right posterior, we must choose \mathbf{F} , \mathbf{W} and σ_ξ such that $\boldsymbol{\mu}^r(\mathbf{h}) = \boldsymbol{\mu}(\mathbf{h})$ and $\Sigma^r = \Sigma$. A possible combination is

$$\mathbf{F} = (\sigma_\xi/\sigma_h)^2 \mathbf{A}^\top \quad (10)$$

$$\mathbf{W} = \mathbf{W}_L \equiv \mathbf{I} - \sigma_\xi^2 \Sigma^{-1} \quad (11)$$

$$\sigma_\xi \text{ arbitrary, } > 0 \quad (12)$$

In the study that follows, we will be interested in the likelihood matrix \mathbf{A} only insofar as it affects the posterior covariance matrix Σ . We can in fact ignore \mathbf{A} altogether, and focus on the case where $\mathbf{h} = 0$, so that the posterior collapses to the prior with a covariance matrix $\Sigma = \mathbf{C}$ whose structure turns out to be the only thing that affects the speed of sampling.

3 Langevin sampling is very slow

Langevin sampling [4, 10, 11] is a common sampling technique, in which a stochastic dynamical system performs a “noisy gradient ascent of the log posterior”:

$$d\mathbf{r} = \frac{\partial}{\partial \mathbf{r}} \log p(\mathbf{r}|\mathbf{h}) d\mathbf{t} + d\boldsymbol{\xi} \quad (13)$$

(where $d\boldsymbol{\xi}$ is a unitary Wiener process). When $\mathbf{r}|\mathbf{h}$ is Gaussian, Equation (13) reduces to Equation (7) for $\sigma_\xi = 1$ and the choice of \mathbf{F} and \mathbf{W} given in Equations (10) and (11) – hence the notation \mathbf{W}_L . Note that $\mathbf{W}^L = \mathbf{W}_L^\top$, i.e. it is a symmetric weight matrix.

As we show now, this choice of weight matrix leads to very slow mixing (i.e. very long correlation times for $\mathbf{r}(t)$) in any high-dimensional sampling space ($N \gg 1$). In a linear network, the average autocorrelation length is dominated by the decay time constant τ_{\max} of the slowest eigenmode, i.e. the eigenvector of $(\mathbf{W} - \mathbf{I})$ associated with the eigenvalue $\lambda_{\max}^{\mathbf{W}-\mathbf{I}}$ which, of all the eigenvalues of $(\mathbf{W} - \mathbf{I})$, has the largest real part (which must still be negative, for stability reasons). The contribution of that slowest eigenmode to the sample autocorrelation time is roughly $\tau_{\max} = -\tau_m / \text{Re}(\lambda_{\max}^{\mathbf{W}-\mathbf{I}})$, so sampling becomes very slow when $\text{Re}(\lambda_{\max}^{\mathbf{W}-\mathbf{I}})$ approaches 0 (from below). This is, in fact, what happens with Langevin sampling as $N \rightarrow \infty$. To see this, let us recall that $(\mathbf{W}_L - \mathbf{I})$ is real and symmetric, so its eigenvalues are all real, and since

$\mathbf{W}_L - \mathbf{I} = -\sigma_\xi^2 \Sigma^{-1}$ we can write¹

$$\lambda_{\max}^{\mathbf{W}_L - \mathbf{I}} = -\sigma_\xi^2 \lambda_{\min}^{\Sigma^{-1}} = -\frac{\sigma_\xi^2}{\lambda_{\max}^\Sigma} \quad (14)$$

Now, again because of its symmetry, Σ is a normal matrix, and so it is similar to (or equal to the unitary transformation of) a diagonal matrix that contains its eigenvalues. Since unitary transformations preserve the Frobenius norm, we can write

$$\sum_{i,j} \Sigma_{ij}^2 = \sum_i (\lambda_i^\Sigma)^2 \quad (15)$$

and since all the eigenvalues of Σ are positive,

$$N (\lambda_{\max}^\Sigma)^2 \geq \sum_i (\lambda_i^\Sigma)^2 \quad (16)$$

Combining Eqs. 14-16, we arrive at a bound that relates the maximum eigenvalue of $(\mathbf{W}_L - \mathbf{I})$ to a basic summary statistics of the posterior covariance matrix Σ :

$$\lambda_{\max}^{\mathbf{W}_L - \mathbf{I}} \geq -\sigma_\xi^2 \sqrt{\frac{N}{\sum_{i,j} \Sigma_{ij}^2}} \quad (17)$$

In the $N \rightarrow \infty$ limit, assuming that pairwise correlations do not vanish, the denominator is expected to be $\mathcal{O}(N^2)$, meaning that $0 > \lambda_{\max}^{\mathbf{W}_L - \mathbf{I}} \geq -\mathcal{O}(1/\sqrt{N})$: the slowest eigenmode of \mathbf{W}_L becomes critically slow for high-dimensional posteriors. To formalize this intuition, let us assume that $\Sigma_{ii} \simeq \sigma_0^2$ (all posterior variances are roughly equal) and that the distribution of pairwise posterior correlations has zero mean and standard deviation σ_r . We can then rewrite Equation (17) as

$$\lambda_{\max}^{\mathbf{W}_L - \mathbf{I}} \geq \frac{-(\sigma_\xi/\sigma_0)^2}{\sqrt{1 + N\sigma_r^2}} \quad (18)$$

We see that Langevin sampling is bound to be slow in the limit of large state spaces ($N \rightarrow \infty$) and when pairwise correlations do not vanish² in that limit (Figure 3, dashed lines).

We can refine this bound in the case when Σ is drawn from a Wishart distribution with n degrees of freedom and scale matrix $(\sigma_0^2/n)\mathbf{I}$ (Figure 3). In this case, the expected value of a diagonal element (variance) in Σ is σ_0^2 , and the distribution of pairwise correlations is centered with variance $\sigma_r^2 \approx 1/n$. If $\sigma_r^2 \sim \mathcal{O}(1)$, Σ becomes low-rank as N grows, and in fact, it has only $\approx \sigma_r^{-2}$ non-zero eigenvalues³. Equation (16) can be adjusted to take this

¹For a non-singular matrix \mathbf{M} , the eigenvalues of \mathbf{M}^{-1} are the inverses of those of \mathbf{M} ; and since Σ is a positive definite covariance matrix, all its eigenvalues are positive, which yields Equation (14).

²Or vanish, but not as fast as $1/\sqrt{N}$.

³In principle, a singular Σ would not be an appropriate posterior covariance matrix – and indeed, no linear stochastic network such as described by Equation (7) would be able to sample from $\mathcal{N}(\boldsymbol{\mu}, \Sigma)$ then. Introducing a small regularizer, i.e. considering $\tilde{\Sigma} \equiv \Sigma + \varepsilon^2 \mathbf{I}$, solves the problem but does not alter the asymptotic properties of the bound we derive here, so we omit this detail for the sake of clarity.

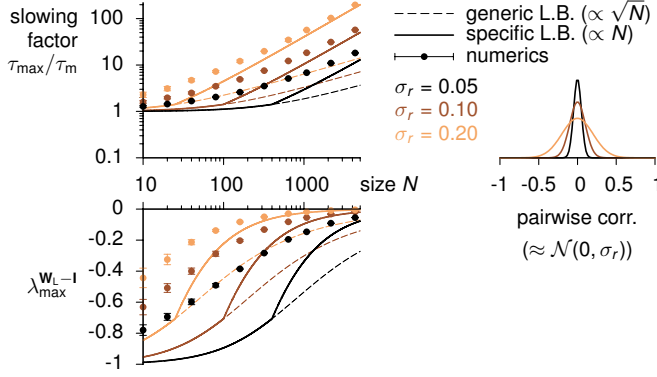


Figure 3: Langevin sampling is slow in high-dimension. Random covariance matrices Σ of size N are drawn from a Wishart distribution with n degrees of freedom and scale matrix $\sigma_0^2 \mathbf{I}/n$. This yields an average variance of σ_0^2 , and a distribution of pairwise correlations with zero mean and variance $\sigma_r^2 \approx 1/n$ (right). Sampling from $\mathcal{N}(\cdot, \Sigma)$ using a stochastic neural network (cf. Figure 2) with $\mathbf{W} = \mathbf{W}_L$ (Langevin, symmetric solution) becomes increasingly slow as N grows, as indicated by the relative decay time constant τ_{\max}/τ_m of the slowest eigenmode of $(\mathbf{W}_L - \mathbf{I})$ (top), which is related to the inverse of its largest eigenvalue (bottom). Dots indicate the numerical evaluation of the corresponding quantities, for 100 sample matrices for each N . Dashed lines correspond to the generic bound in Equation (18). Solid lines are refined bounds for the specific case $\Sigma \sim \text{Wishart}(\sigma_0^2 \mathbf{I}/n, n)$ with $n \approx 1/\sigma_r^2$ (Equation (21)). The two bounds merge for $N < n$. Parameters were set to $\sigma_\xi = \sigma_0 = 1$.

into account:

$$\min(n, N) (\lambda_{\max}^\Sigma)^2 \geq \sum_i (\lambda_i^\Sigma)^2 \quad (19)$$

As mentioned above, the r.h.s. of Equation (19) is equal to the squared Frobenius norm of Σ , which can be easily estimated for the Wishart ensemble, at least for N and n not too small:

$$\|\Sigma\|_F^2 \approx N \sigma_0^2 \left(1 + \frac{N}{n}\right) \quad (20)$$

Using Equation (14) and recalling that $n \approx \sigma_r^{-2}$, we arrive at a Wishart-specific bound:

$$\lambda_{\max}^{\mathbf{W}_L - \mathbf{I}} \geq \frac{-(\sigma_\xi/\sigma_0)^2 \sqrt{\min(\sigma_r^{-2}, N)}}{\sqrt{N(1 + \sigma_r^2 N)}} \quad (21)$$

Note that when $\sigma_r < 1/\sqrt{N}$, the bound becomes equivalent to the more general case in Equation (18). For $\sigma_r \sim \mathcal{O}(1)$, however, the slowing problem becomes worse for Wishart matrices, since now $0 > \lambda_{\max}^{\mathbf{W}_L - \mathbf{I}} \geq -\mathcal{O}(1/N)$ (Figure 3, solid lines).

The ratio (σ_0/σ_ξ) , which shows up in both versions of our bound (Equations (18) and (21)), tells us how much the

recurrent interactions must amplify the external noise in order to produce from the right stationary activity distribution (recall that σ_ξ measures the magnitude of the activity fluctuations due to the input noise alone, in the absence of recurrent circuitry). The more amplification is required ($\sigma_\xi \ll \sigma_0$), the slower the dynamics of Langevin sampling is bound to become.

In summary, Langevin sampling corresponds to symmetric interactions (which violates Dale’s law), and in the Gaussian case considered here, yields unacceptably slow mixing in high-dimensional latent spaces. This should be true whenever i) the magnitude of the posterior variances does not depend on N , and ii) the spread of the distribution of posterior pairwise correlations also does not depend on N . The types of generative models under which the second assumption holds are yet to be characterized; we leave this to future work.

4 What is the fastest sampler?

While Langevin dynamics (Equation (13)) give a general recipe for sampling from any given posterior density, it unduly constrains the dynamics to obey symmetric interactions – at least in the Gaussian case. To see why this is a huge restriction, let us observe that *any connectivity matrix of the form*

$$\mathbf{W}(\mathbf{S}) = \mathbf{I} + (-\sigma_\xi^2 \mathbf{I} + \mathbf{S}) \Sigma^{-1} \quad (22)$$

where \mathbf{S} is an arbitrary skew-symmetric matrix, solves Equation (9), and therefore induces the right stationary distribution $\mathcal{N}(\cdot, \Sigma)$ under the linear stochastic dynamics of Equation (7). Note that Langevin sampling (Equation (11)) corresponds to $\mathbf{S} = 0$. But in general, there are $\mathcal{O}(N^2)$ degrees of freedom in the skew-symmetric matrix \mathbf{S} , which is a lot. Could these be exploited to speed up mixing? In the following, we show that indeed a large gain in sampling speed can be obtained through an appropriate choice of \mathbf{S} . By formulating the problem as one of robust control, we can use optimization techniques to discover the *fastest sampler*, in a sense that we define below.

Let $\mathbf{K}(\mathbf{S}, \tau) = \langle \delta \mathbf{r}(t) \delta \mathbf{r}(t + \tau)^\top \rangle_t$ be the covariance matrix between pairs of samples separated by a time interval τ , in the stationary regime (we use the notation $\delta \mathbf{r}(t) \equiv \mathbf{r}(t) - \boldsymbol{\mu}$). Note that $\mathbf{K}(\mathbf{S}, 0) = \Sigma$ is the posterior correlation matrix, and that for fixed σ_ξ^2 and τ_m , $\mathbf{K}(\mathbf{S}, \tau)$ depends only on the interval τ and on the matrix of recurrent weights \mathbf{W} , which itself depends only on our skew-symmetric free-parameter matrix \mathbf{S} . We define a “total slowing cost”

$$\psi_{\text{slow}}(\mathbf{S}) = \frac{1}{2\tau_m N^2} \int_0^\infty \|\mathbf{K}(\mathbf{S}, \tau)\|_F^2 d\tau \quad (23)$$

which penalizes large autocorrelations and pairwise cross-correlations (both positive and negative) in the sequence

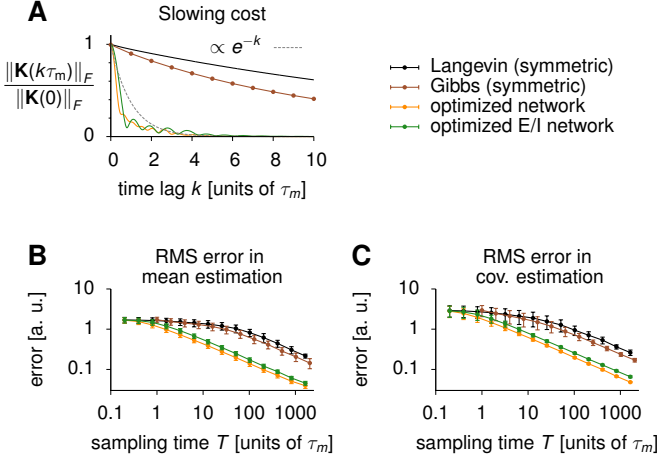


Figure 4: **How fast is the fastest sampler?** (A) Scalar measure of the statistical dependency between any two samples collected τ seconds apart (cf. main text), for Langevin sampling (black), Gibbs sampling (brown, assuming a full update sweep is done every τ_m), the unconstrained optimized network (orange), and the optimized E/I network (green). The gray line shows the behaviour of a purely feedforward network that would merely integrate uncorrelated noise sources with weights given by the Cholesky factor of the target Σ , and decay with a time constant τ_m . Optimized recurrent networks do better, because they can evolve *collectively* faster than an isolated neuron would. (B) Root-mean-squared error in the estimation of the posterior mean μ on the basis of a limited amount of samples $\mathbf{r}(t)$, collected every 2 ms during a period T (x-axis, in units of τ_m). Errorbars denote standard deviation over many trials. (C) Same as in (B), for posterior variance/covariance estimation.

of samples generated by the circuit dynamics. Here $\|\mathbf{M}\|_F^2 \equiv \text{trace}(\mathbf{M}\mathbf{M}^\top) = \sum_{ij} M_{ij}^2$ is the squared Frobenius norm of \mathbf{M} .

We can use this measure of slowness to illustrate the slow mixing behaviour of Langevin sampling on a toy covariance matrix. We generate a random Σ of size $N = 100$ from a full, random orthonormal basis $(\mathbf{u}_1, \dots, \mathbf{u}_N)$ ⁴ as:

$$\Sigma = \frac{\sigma_0^2 - (1-p)}{p} \sum_{i=1}^{pN} \mathbf{u}_i \mathbf{u}_i^\top + \sum_{i=pN+1}^N \mathbf{u}_i \mathbf{u}_i^\top \quad (24)$$

One can easily check that the average variance, i.e. $\text{trace}(\Sigma)/N$, is equal to σ_0^2 , which we set to 3. We choose $p = 0.1$, resulting in a fairly broad distribution of pairwise correlation coefficients in Σ ($\sigma_r \approx 0.15$). Figure 4A illustrates the behaviour of Langevin sampling by plotting $\|\mathbf{K}(\mathbf{S} = 0, \tau)\|_F$ as a function of the time lag τ : as predicted above in Section 3, mixing is indeed an order of magnitude slower than the single-neuron time constant

⁴obtained by Gram-Schmidt orthonormalization of a set of N random Gaussian vectors.

τ_m . Note that ψ_{slow} in Equation (23) is proportional to the area under the squared curve in Figure 4A. The slow dynamics of Langevin sampling are also illustrated in Figure 7B (top), in which 500 ms of network activity are shown.

Using the same measure, we can also look at the speed of Gibbs sampling, another widely used sampling technique (e.g. [12, 13]). It is defined as a Markov chain that operates in discrete time, and to compare its mixing speed with that of our linear stochastic dynamics, we assume that a single discrete step (in which all neurons have been updated once) consumes a time τ_m . The speed of Gibbs sampling is comparable to that of Langevin here: samples are still very correlated on a timescale of order $\sim 10 \tau_m$.

We now show that the skew-symmetric matrix \mathbf{S} can be optimized for sampling speed, by directly minimizing the slowing cost $\psi_{\text{slow}}(\mathbf{S})$, subject to an L_2 -norm penalty. Our overall cost function is therefore:

$$\mathcal{L}(\mathbf{S}) \equiv \psi_{\text{slow}}(\mathbf{S}) + \frac{\lambda_{L_2}}{2N^2} \|\mathbf{W}(\mathbf{S})\|_F^2 \quad (25)$$

It is well known [9] that $\mathbf{K}(\mathbf{S}, \tau)$ obeys the following differential equation:

$$\tau_m \frac{d\mathbf{K}(\mathbf{S}, \tau)}{d\tau} = [\mathbf{W}(\mathbf{S}) - \mathbf{I}] \mathbf{K}(\mathbf{S}, \tau) \quad (26)$$

such that for $\tau \geq 0$

$$\mathbf{K}(\mathbf{S}, \tau) = e^{[\mathbf{W}(\mathbf{S}) - \mathbf{I}] \tau / \tau_m} \Sigma \quad (27)$$

We may thus rewrite $\psi_{\text{slow}}(\mathbf{S})$ as

$$\psi_{\text{slow}}(\mathbf{S}) = \frac{1}{2N^2} \text{tr} \left[\int_0^\infty e^{\tau[\mathbf{W}(\mathbf{S}) - \mathbf{I}]} \Sigma^2 e^{\tau[\mathbf{W}(\mathbf{S})^\top - \mathbf{I}]} d\tau \right] \quad (28)$$

The derivatives w.r.t \mathbf{W} are given by [14]:

$$\frac{\partial \psi_{\text{slow}}(\mathbf{S})}{\partial \mathbf{W}} = \frac{\mathbf{Q}\mathbf{P}}{N^2} \quad (29)$$

where matrices \mathbf{P} and \mathbf{Q} are the solutions of the following Lyapunov equation pair:

$$(\mathbf{W} - \mathbf{I})\mathbf{P} + \mathbf{P}(\mathbf{W} - \mathbf{I})^\top = -\Sigma^2 \quad (30)$$

$$(\mathbf{W} - \mathbf{I})^\top \mathbf{Q} + \mathbf{Q}(\mathbf{W} - \mathbf{I}) = -\mathbf{I} \quad (31)$$

These equations can be solved efficiently [15], e.g. using the Matlab function `lyap`. Note also that $\psi_{\text{slow}}(\mathbf{S}) = \text{tr}(\mathbf{P})/2N^2$ [14]. Now, a straightforward application of the chain rule yields

$$\frac{\partial \psi_{\text{slow}}(\mathbf{S})}{\partial \mathbf{S}} = \frac{1}{N^2} [(\Sigma^{-1}\mathbf{P}\mathbf{Q})^\top - (\Sigma^{-1}\mathbf{P}\mathbf{Q})] \quad (32)$$

which is skew-symmetric, as it should. The L_2 -penalty term in Equation (25) is more easily differentiated, yielding an overall gradient

$$\frac{\partial \mathcal{L}(\mathbf{S})}{\partial \mathbf{S}} = \frac{1}{N^2} [(\Sigma^{-1}\mathbf{P}\mathbf{Q})^\top - (\Sigma^{-1}\mathbf{P}\mathbf{Q})] + \frac{\lambda_{L_2}}{N^2} [\mathbf{S}\Sigma^{-2} + \Sigma^{-2}\mathbf{S}] \quad (33)$$

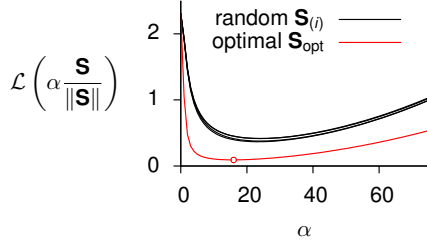


Figure 5: The cost function \mathcal{L} (Equation (25)) evaluated along 4 different random directions $\mathbf{S}_{(i)}$ in the space of skew-symmetric matrices (black) and along the optimal direction \mathbf{S}_{opt} (red) which we found by gradient descent. The red dot indicates the minimum of \mathcal{L} .

This gradient may be used in any gradient-based optimization approach to minimize $\mathcal{L}(\mathbf{S})$ and obtain the fastest regularized sampler, which we now show on the toy covariance matrix of Equation (24).

We initialized \mathbf{S} with random, weak and uncorrelated elements ($S_{i>j} \sim \mathcal{N}(0, 0.001^2)$, $S_{ji} = -S_{ij}$ and $S_{ii} = 0$), and ran the nonlinear conjugate gradient algorithm (golden section line search and Polak-Ribière approximation) to minimize the regularized cost function in Equation (25) with $\lambda_{L_2} = 0.1$. The final, optimal \mathbf{S}_{opt} induces a weight matrix \mathbf{W}_{opt} given by Equation (22) and shown in Figure 7A (center). Importantly, \mathbf{W}_{opt} is no longer symmetric, and its elements are an order of magnitude larger than in the Langevin symmetric solution \mathbf{W}_L . Note also that the cost function seems to be convex along any random direction in the space of skew-symmetric matrices (Figure 5), suggesting (but not proving⁵) that $\mathcal{L}(\mathbf{S})$ has a single minimum, and therefore that the matrix \mathbf{S}_{opt} corresponds to the fastest sampler.

The optimal sampler is an order of magnitude faster than either Langevin or Gibbs sampling: samples are decorrelated on a timescale that is even faster than the single-neuron time constant τ_m (Figure 4A, orange). Such decorrelation dramatically improves the sample-based estimation of the posterior mean and covariances, as shown in Figure 4B and C. For any sampler that has the right stationary distribution, the difference between the sample estimates (samples collected every 5 ms) of $\boldsymbol{\mu}$ and $\boldsymbol{\Sigma}$ and their true values vanishes with sampling time, *asymptotically*. For a finite number of samples, however, the estimation error depends on how independent those samples are. For both Langevin and Gibbs sampling, the RMS error in parameter estimation starts decaying as $1/\sqrt{T}$ (the expected asymptotic decay rate) only after $T > 10$ seconds of sampling time. In contrast, the asymptotic rate is reached by the optimal sampler after only $\tau_m = 20$ ms, that is, from the very first sample.

We note in passing from the result of Figure 5 that

⁵For example, there could be multiple local minima hiding on a sphere $\|\mathbf{S}\| = \text{constant}$.

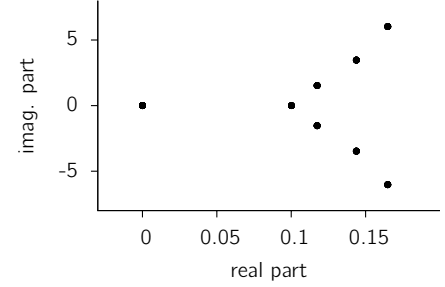


Figure 6: **Eigenvalue spectrum of the connectivity matrix of the fastest sampler.** This spectrum is extremely structured: each “dot” that forms the “rotated V” on the right is actually made of 10 almost-identical eigenvalues, while the left-most “dot” is made of 30 zero eigenvalues.

a decent (though sub-optimal) sampling speed can be achieved without fine-tuning, through the use of a *random* skew symmetric matrix \mathbf{S} (cf. black curves).

Perhaps intriguingly, the eigenvalue spectrum of \mathbf{W}_{opt} is highly structured (Figure 6). Moreover, the sum of its eigenvalue squared moduli accounts for only 25% of $\|\mathbf{W}_{\text{opt}}\|_F^2$, indicating \mathbf{W}_{opt} is strongly nonnormal⁶ [7]; Indeed, for a normal matrix \mathbf{W} – such as the Langevin solution \mathbf{W}_L –, one would expect $\sum_i |\lambda_i|^2 = \|\mathbf{W}\|_F^2$. Deviation from normality can have important consequences for the dynamics in a linear stochastic network such as the one we consider here: the eigenvectors of \mathbf{W} are expected *not* to be orthogonal, such that the apparent activity decay along those eigenvectors (at a speed governed by the corresponding eigenvalue real part) can hide large albeit transient amplification of momentary perturbations along some other directions in state space [8, 16, 17, 18]. It is illuminating to visualize activity trajectories in the plane defined by the topmost and bottommost eigenvectors of $\boldsymbol{\Sigma}$, i.e. the first and last principal components (PCs) of the network activity (Figure 7C). Inspecting these trajectories in 5 ms time steps, we see that the distribution of discrete increments generated by Langevin sampling are identical in both directions. Since accurate sampling requires the last PC to have small variance (at least relative to the first PC), those distributions of increments must be narrow, which explains the slowness of Langevin sampling: a lot of very small steps must be taken along the first PC. In contrast, the optimal network is not limited by the last PC, and can indeed make much larger transient excursions along the first PC (Figure 7C, middle).

⁶“Nonnormal” here has nothing to do with “non-Gaussian”: \mathbf{M} is nonnormal iff it is *not normal*, i.e. $\mathbf{M}\mathbf{M}^\top \neq \mathbf{M}^\top\mathbf{M}$.

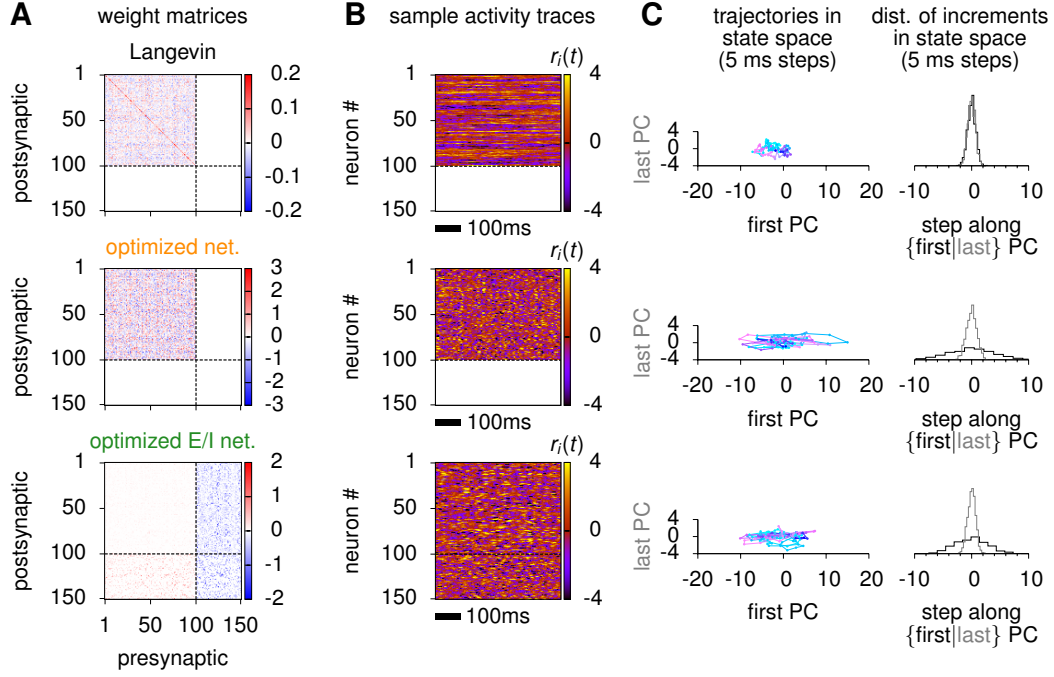


Figure 7: **Fast sampling with optimized networks.** (A) Synaptic weight matrices for the Langevin network (top), the fastest sampler (middle) and the fastest sampler that obeys Dale’s law (bottom). Note that the synaptic weights in both optimized networks are an order of magnitude larger than in the symmetric Langevin solution. The first two networks are of size $N = 100$, so the last 50 columns and rows are empty here, but reproduced for comparison with the optimized E/I network which has size $N = 150$ (bottom). (B) 500 ms of spontaneous network activity ($\mathbf{h} = 0$) for each of the three networks, for all of which the stationary distribution of \mathbf{r} (restricted to the first 100 neurons) is the same multivariate Gaussian. (C) Left: activity trajectories (the same 500 ms as shown in (B)) in the plane defined by the topmost and bottommost eigenvectors of the posterior covariance matrix Σ (corresponding to the first and last principal components of the activity fluctuations $\mathbf{r}(t)$). For the E/I network, the projection is restricted to the excitatory neurons. Right: distribution of increments along both axes, measured in 5 ms time steps. Langevin sampling takes steps of equal size along all directions, while the optimized networks take much larger steps along the directions of large variance prescribed by the posterior.

5 Balanced E/I networks for fast sampling

We now consider more plausible network structures, namely balanced networks made of neurons that are either excitatory or inhibitory (Dale’s law). We assume that there are $N_{\text{exc.}} = N$ excitatory neurons, where N is the dimension of the distribution we want to sample from, and $N_{\text{inh.}}$ inhibitory neurons whose activity distribution is irrelevant (i.e. we regard inhibitory neurons as auxiliary sampling variables, in the spirit of Hamiltonian Monte Carlo methods [10]). In the following, we set $N_{\text{exc.}} = 100$ and $N_{\text{inh.}} = 50$. Let $M = N_{\text{exc.}} + N_{\text{inh.}}$ denote the total network size. We assume similar stochastic dynamics as before, i.e.

$$d\mathbf{r} = \frac{dt}{\tau_m} (-\mathbf{r}(t) + \mathbf{W}\mathbf{r}(t) + \mathbf{F}\mathbf{h}) + \chi d\boldsymbol{\xi}(t) \quad (34)$$

where χ is a diagonal matrix of cell type-specific input noise variances:

$$\chi_{ii} = \begin{cases} \chi_{\text{exc.}} & \text{if } i \leq N_e \\ \chi_{\text{inh.}} & \text{otherwise.} \end{cases} \quad (35)$$

Here $\chi_{\text{exc.}}$ and $\chi_{\text{inh.}}$ are two free parameters.

The connectivity matrix \mathbf{W} is now made of $N_{\text{exc.}}$ positive columns followed by $N_{\text{inh.}}$ negative columns. This makes it difficult to apply the above approach (Section 4) to find the fastest E/I sampler, as picking an arbitrary skew-symmetric matrix \mathbf{S} in Equation (22) will not yield the column sign structure of an E/I network in general. Therefore, we no longer have a parametric form for the solution matrix manifold on which to find the fastest network. However, with a few simple variations, we can still formulate the problem as one of unconstrained optimization, as explained now.

The first step is to parameterize \mathbf{W} as follows:

$$W_{ij} = \delta_{ij} s_j \exp \beta_{ij} \quad (36)$$

where s_j is a fixed sign that depends only on presynaptic neuron j ($s_j = +1$ for $j \neq N_{\text{exc.}}$, -1 otherwise), and the β_{ij} 's are free parameters. Note that we do not allow for autapses (δ_{ij} term in Equation (36)). Second, since the target posterior distribution specifies only the $N_{\text{exc.}} \times N_{\text{exc.}}$ upper-left quadrant Σ of the overall covariance matrix which we denote by Λ , we are free to optimize over the other quadrants. We parameterize Λ by its Cholesky factor:

$$\Lambda = \mathbf{L}\mathbf{L}^\top, \quad \mathbf{L} \equiv \begin{pmatrix} \mathbf{L}_{11} & 0 \\ \mathbf{L}_{12} & \mathbf{L}_{22} \end{pmatrix} \quad (37)$$

where \mathbf{L}_{11} is the Cholesky factor of the posterior covariance matrix Σ (i.e. $\Sigma = \mathbf{L}_{11}\mathbf{L}_{11}^\top$), and the two matrices \mathbf{L}_{12} and \mathbf{L}_{22} are free parameters. Note that \mathbf{L}_{12} is a full rectangular matrix of size $N_{\text{inh.}} \times N_{\text{exc.}}$, while \mathbf{L}_{22} is lower-triangular with dimensions $N_{\text{inh.}} \times N_{\text{inh.}}$. Third, we incorporate the Lyapunov equation (Equation (9)) as an additional constraint in our objective function, which becomes

$$\mathcal{L}(\theta) = \psi_\Lambda(\theta) + \lambda_{\text{slow}} \psi_{\text{slow}}(\theta) + \lambda_{L_2} \|\mathbf{W}\|_F^2 \quad (38)$$

with

$$\psi_\Lambda(\theta) = \frac{1}{2M^2} \|(\mathbf{W} - \mathbf{I})\Lambda + \Lambda(\mathbf{W} - \mathbf{I})^\top + \tau_m \mathbf{X}^2\|_F^2 \quad (39)$$

and ψ_{slow} is defined as in Equation (23) with $\mathbf{K}(\theta, \tau) \equiv \langle \delta \mathbf{r}_e(t) \delta \mathbf{r}_e(t + \tau)^\top \rangle$. Here $\mathbf{r}_e(t)$ is the vector of network activity \mathbf{r} at time t , in which the inhibitory neurons' firing rates have been replaced by zeros, i.e. $\mathbf{r}_e(t) = \mathbf{J}\mathbf{r}(t)$ with

$$\mathbf{J} = \begin{pmatrix} \mathbf{I} & 0 \\ 0 & 0 \end{pmatrix}. \quad (40)$$

When the cost term ψ_Λ is zero, then the Lyapunov equation

$$(\mathbf{W} - \mathbf{I})\Lambda + \Lambda(\mathbf{W} - \mathbf{I})^\top = -\tau_m \mathbf{X}^2 \quad (41)$$

is satisfied, and therefore Λ is the stationary covariance matrix of the network activity. In particular, Σ is the covariance matrix of the excitatory neurons' activity, as wanted.

Finally, the vector θ comprises all the free parameters we have in this problem, i.e. the private noise variances $\chi_{\text{exc.}}$ and $\chi_{\text{inh.}}$, all synaptic weight parameters β_{ij} , and all the relevant elements of \mathbf{L}_{12} and \mathbf{L}_{22} .

The gradients of $\psi_{\text{slow}}(\theta)$ can be obtained as in Section 4 (cf. also [14]):

$$\psi_{\text{slow}} = \frac{1}{2N^2} \text{tr}(\mathbf{J}\Sigma\mathbf{Q}\Sigma) \quad (42)$$

$$\frac{\partial \psi_{\text{slow}}}{\partial \mathbf{W}} = \frac{\mathbf{Q}\mathbf{P}}{N^2} \quad (43)$$

$$\frac{\partial \psi_{\text{slow}}}{\partial \mathbf{L}} = \frac{1}{N^2} [\mathbf{J}\Lambda\mathbf{Q} + (\mathbf{J}\Lambda\mathbf{Q})^\top] \mathbf{L} \quad (44)$$

where \mathbf{P} and \mathbf{Q} solve

$$(\mathbf{W} - \mathbf{I})\mathbf{P} + \mathbf{P}(\mathbf{W} - \mathbf{I})^\top = -\Lambda\mathbf{J}\Lambda \quad (45)$$

$$(\mathbf{W} - \mathbf{I})^\top \mathbf{Q} + \mathbf{Q}(\mathbf{W} - \mathbf{I}) = -\mathbf{J} \quad (46)$$

The derivatives w.r.t $\chi_{\text{exc.}}$ and $\chi_{\text{inh.}}$, as well as the gradients of the other cost terms (ψ_Λ and $\|\mathbf{W}\|_F^2$), are more easily derived.

We performed nonlinear conjugate gradients to minimize the cost function $\mathcal{L}(\theta)$ in Equation (38) with $\lambda_{L_2} = \lambda_{\text{slow}} = 0.1$. The results are presented in a similar format as before, in the same figures (green lines). The resulting synaptic weight matrix is shown in Figure 7A (bottom), together with a 500 ms-long activity sample (Figure 7B, bottom). This Dale-compliant solution is almost as fast as the best (regularized) unconstrained network (compare orange and green in Figure 4), indicating that Dale's law – unlike the symmetry constraint implicitly present in Langevin sampling – is not fundamentally detrimental to mixing speed.

6 Discussion

We have studied sampling for Bayesian inference in neural circuits, and observed that a linear stochastic network is able to sample from the posterior under a linear Gaussian latent variable model. Hidden variables are directly encoded in the activity of single neurons, and their joint activity undergoes moment-to-moment fluctuations and visits each portion of latent state space with a frequency that matches the corresponding, prescribed posterior density. To achieve this, external noise sources fed into the network are amplified by the recurrent circuitry, but preferentially amplified along the state-space directions that matter.

We have shown that a popular machine learning technique, namely Langevin sampling [4, 10, 11], can be mapped onto such neuronal network dynamics with what turns out to be an unfortunate choice of a *symmetric weight matrix*. There, an analytical argument predicts dramatic slowing in high-dimensional latent spaces, also consistent with numerical simulations. Samples are correlated on a timescale that extends much beyond the single-neuron decay time constant.

When the above symmetry constraint is relaxed, a family of other solutions opens up that can potentially lead to much faster sampling. We chose to explore this possibility from a normative viewpoint, and optimized the network connectivity directly for speed of sampling. The fastest sampler turned out to be very asymmetric, non-normal in the mathematical sense, and typically an order of magnitude faster than Langevin sampling.

Notably, we also found that constraining each neuron to be either excitatory or inhibitory, but not of a mixed type,

does not impair the performance of the fastest sampler but bridges the gap of biological plausibility.

Our fast samplers are capable of taking large steps in directions of large posterior variance, and small steps in other directions. This, together with the interpretation of Langevin sampling as a stochastic gradient ascent on the log-posterior, suggests a link between our optimal sampling scheme and natural gradient algorithms [19], and potentially also with Riemannian Monte Carlo sampling [20].

Acknowledgments

This work was supported by a fellowship from the Swiss National Science Foundation (G. H.), the Wellcome Trust (M. L. and G. H.) and the Gatsby Charitable Foundation (L. A.).

References

- [1] M. Lengyel, G. Hennequin, and L. Aitchison. Fast sampling in recurrent neural circuits. In *Cosyne abstracts 2014*, Salt Lake City, USA, 2014.
- [2] S. Thorpe, D. Fize, and C. Marlot. Speed of processing in the human visual system. *Nature*, 381:520–522, 1996.
- [3] D. Knill and A. Pouget. The bayesian brain: the role of uncertainty in neural coding and computation. *Trends in Neurosciences*, 27:712–719, 2004.
- [4] D. MacKay. *Information theory, inference, and learning algorithms*. Cambridge University Press, 2003.
- [5] J. Fiser, P. Berkes, G. Orbán, and M. Lengyel. Statistically optimal perception and learning: from behavior to neural representations. *Trends in Cognitive Sciences*, 14:119–130, 2010.
- [6] P. Berkes, G. Orbán, M. Lengyel, and J. Fiser. Spontaneous cortical activity reveals hallmarks of an optimal internal model of the environment. *Science*, 331:83–87, 2011.
- [7] L. N. Trefethen and M. Embree. *Spectra and pseudospectra: the behavior of nonnormal matrices and operators*. Princeton University Press, 2005.
- [8] B. K. Murphy and K. D. Miller. Balanced amplification: A new mechanism of selective amplification of neural activity patterns. *Neuron*, 61:635–648, 2009.
- [9] C. W. Gardiner. *Handbook of stochastic methods: for physics, chemistry, and the natural sciences*. Berlin: Springer, 1985.
- [10] R. Neal. MCMC using hamiltonian dynamics. *Handbook of Markov Chain Monte Carlo*, page 113162, 2011.
- [11] M. Welling and Y. W. Teh. Bayesian learning via stochastic gradient Langevin dynamics. In *Proceedings of the International Conference on Machine Learning*, 2011.
- [12] M. Mezard and A. Montanari. *Information, physics, and computation*. Oxford University Press, 2009.
- [13] L. Buesing, J. Bill, B. Nessler, and W. Maass. Neural dynamics as sampling: a model for stochastic computation in recurrent networks of spiking neurons. *PLoS Computational Biology*, 7:1–22, 2011.
- [14] J. Vanbiervliet, B. Vandereycken, W. Michiels, S. Vandewalle, and M. Diehl. The smoothed spectral abscissa for robust stability optimization. *SIAM J on Optimization*, 20:156171, 2009.
- [15] R. H. Bartels and G. W. Stewart. Solution of the matrix equation $AX+XB=C$. *Communications of the ACM*, 15:820–826, 1972.
- [16] Mark S. Goldman. Memory without feedback in a neural network. *Neuron*, 61:621–634, 2009.
- [17] G. Hennequin, T. P. Vogels, and W. Gerstner. Non-normal amplification in random balanced neuronal networks. *Physical Review E*, 86:011909, 2012.
- [18] G. Hennequin, T. P. Vogels, and W. Gerstner. Optimal control of transient dynamics in balanced networks supports generation of complex movements. *Neuron*, accepted, 2014.
- [19] S. I. Amari. Natural gradient works efficiently in learning. *Neural computation*, 10:251276, 1998.
- [20] M. Girolami and B. Calderhead. Riemann manifold langevin and hamiltonian monte carlo methods. *Journal of the Royal Statistical Society: Series B (Statistical Methodology)*, 73(2):123214, 2011.